Making public use, synthetic files of longitudinal establishment data

Jerry Reiter
Duke University

Satkartar Kinney
Duke University

August 2006

Longitudinal business data are widely desired by researchers, but difficult to make available to the public because of confidentiality constraints. In this paper, we discuss the generation of synthetic public use datasets for establishment data. The basic idea is to release simulated values of sensitive variables, generated from probability distributions fit using genuine data. This can protect confidentiality, since attributes are synthetic rather than real. And, when the models describe the data well, broad-scale inferences from the synthetic datasets will be similar to those from the genuine data. We illustrate approaches for generating synthetic establishment data by using LEHD infrastructure data.